Integrating Sensory Data for Object Recognition Tasks

Peter K. Allen and Ruzena Bajcsy

Department of Computer Science
Columbia University
New York, New York 10027

Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104

Abstract

Object recognition is a difficult task for single sensor systems (e.g. machine vision) in unconstrained environments. A useful approach is to combine sensory data from more than one source to overcome these problems. However, using multiple sensors poses new problems with respect to coordination of the sensors, strategies for their use and integration of their data. In this paper, these problems are explored and solutions posed for the task of object recognition using passive stereo vision and active tactile sensing.

1. INTRODUCTION

Many robotic tasks are attempted without sensing, assuming an absolute world model that never changes. For example, in many pick and place operations, the objects are always in a previously known absolute position and orientation. This approach offers little flexibility. Robotic systems need the ability to use sensory feedback to understand their environment. Work environments are not static and cannot always be adequately constrained. There is much uncertainty in the world, and we as humans are equipped with powerful sensors to deal with this uncertainty. Robots need to have this ability also. Incorporating sensory feedback into robotic systems allows nondeterminism to creep into the deterministic control of a robot. There is at present much work going on in the area of sensor design for robotics. Range finders, tactile sensors, force/torque sensors, and other sensors are actively being developed. The challenge to the robotic system builder is to incorporate these sensors into a system and to make use of the data provided by them.

Much of the sensor related work in robotics has tried to use a single sensor to determine environmental properties [1, 5, 7, 9, 10, 18, 17, 19, 22]. A common strategy in computer vision is to try to use vision sensing alone to determine shape properties. Many different "shape" operators have been defined by various researchers trying to isolate separate parts of the visual system that produce depth and surface information. Examples of these are shape from texture [13, 3], shape from shading [12], shape from contour [21, 23, 11] and shape from stereo [14, 8]. A potentially promising idea is to use all of these separate shape operators together in a system that will integrate their results. Unfortunately, the operators all have different sets of constraints on the object's structure, reflectance, and illumination. The integration of these many visual operators is still not well understood. A much more promising approach is to supplement the vision information with other sensory inputs that directly measure the properties of shape we desire. The strategy of trying to obtain enough shape information from a single sensor may fail due to the limitations of that sensor as is typically the case with machine vision. If vision sensing can be supplemented with other sensing information that directly measures shape, more robust and error free descriptions of object structure can result. Multiple sensors can be used in a complementary fashion to extract more information from an environment than a single sensor [20, 16].

This paper is an examination of these issues in general and a specific implementation for the case of integrating vision and touch sensing for the task of object recognition. Touch sensing was chosen for a number of reasons. First, in designing a general purpose robot to work in unconstrained environments, touch sensing is a requirement for tasks such as grasping and manipulation. Second, it is a low cost robotic

Acknowledgement: This work was supported in part by: NSF grant MCS 82--07294, Airforce grant AFOSR 82-NM-299, NIH grant 5-RO1-HL-29985-02, DEC Corp., Lord Corp. and IBM Corp.

sensor that is easily included in a robotic system. Third, and most important, it can directly sense the properties of objects we desire, their position and orientation, without regard to visual occlusion.

The objects to be recognized are common kitchen items; mugs, plates, bowls, pitchers, and utensils. The objects are planar as well as volumetric, contain holes and have concave and convex surfaces. These are fairly complex objects which test the modeling and recognition abilities of most existing systems. The objects are homogeneous in color, with no discernible textures. The lack of surface detail on these objects poses serious problems for many visual recognition systems, since there is a lack of potential features that can be used for matching and depth analysis.

The experimental hardware is shown in figure 1. The objects to be recognized are rigidly placed on the worktable and imaged by a pair of CCD cameras. The tactile sensor is mounted on a 6 degree of freedom PUMA 560 manipulator that receives feedback from the tactile sensor. Figure 2 is an overview of the software of the system. It consists of five distinct modules: the control module, the vision module, the tactile module, the model data base and the matcher. All of these are described in detail in [2]. In this paper, we focus on the integration issue and the design decisions posed by using multiple sensors.

2. DESIGN ISSUES IN MULTIPLE SENSING

There are a number of important and difficult issues posed by using multiple sensors. The first is coordination between the different sensing elements. The trend in sensor design is to have each sensor controlled by its own microcomputer system. Therefore integrating multiple sensors becomes a problem in distributed computing as well. The sensors have different response time, bandwidth, resolution and accuracy. Vision sensing is fast and provides large amounts of data as opposed to a tactile sensor that reports contact/noncontact as it moves slowly over a surface. The nature of the data that each sensor provides is also different. Vision provides image space projections while a tactile sensor reports 3-D points of contact and surface normals. The vision sensors are passive while the tactile sensor is active and requires a larger degree of control. Finally, there is the interaction between the sensors themselves. Do they provide redundant, complementary or disparate data? If two sensors can accomplish a task, what is the best strategy for attempting to sense a part of the scene, given criteria such as maximizing throughput, accuracy or speed. Can the two sensors be used intelligently to reliably understand the three dimensional structure of the objects to be recognized?

2.1. ORGANIZATION OF MULTIPLE SENSORS

The organization used here is a hierarchy, where the sensors are each independent entities that communicate through a central control process. This organization works well if the data and sensing processes are very different as is the case with vision and touch. It is important that the higher levels of the hierarchy not be overly concerned with the details at the lower level. In particular, each sensor system should be able to use its own language and data structures to model the world as it sees it without regard to some global data model. In order to maximize throughput (there are severe real time constraints in robotics) the communication between a sensor and the next level in the hierarchy needs to be minimized. The main method of accomplishing this is by compressing and abstracting the data at each level in the hierarchy, which is easily accomplished by the processors in the sensory hierarchy. Fusion of the data takes place at the top levels of the hierarchy after the data has been abstracted.

2.2. STRATEGIES

The strategies used to integrate multi-sensor data are sensor dependent as would be expected. A careful weighing of the sensor's characteristics that include bandwidth, response time, accuracy, resolution and kind of data will determine these strategies. Some helpful guidelines can be established though. First, there is nothing wrong with overwhelming sensing by many devices to verify and support hypotheses about the world. Redundant sensing that builds confidence levels is not wasteful. It is better to be overwhelmingly correct than partially wrong. As Binford has stated in [4]:

> In machine perception, overwhelming verification of a correct hypothesis is typically inexpensive compared to the computation required to get to the correct hypothesis. These factors shift the utility balance toward getting data needed for a highly constrained decision. Very strong, relevant data are available if descriptive mechanisms can abstract them and interpretation mechanisms use them.

Another important point is that strategies change with domains, sensors and tasks. Therefore, a system should be easily modified to support the development of new strategies.

2.3. ABSTRACTING DATA

Video images are available at 60 hz. The amount of data provided by such a sensor will quickly overwhelm a system, particularly when most of the image is of little interest. Higher levels of reasoning need to have primitives that go beyond pixels and point data which are inherently unstable. For sensors that report geometric primitives the abstraction follows a natural path of points yielding curves yielding surfaces yielding volumes. The benefits from abstracting this data is clear: the creation of powerful, rich and stable entities which can form a basis for high level reasoning about an object in the scene and the reduction of bandwidth between levels.

3. OBJECT RECOGNITION TASKS

Model based object recognition is the paradigm being used to allow higher level knowledge about the domain to be encoded and assist the recognition process. Recognition has two components, a data driven or bottom up component that supplies low level feature and primitive information and a high level that utilizes these primitives to understand a scene. At some point, low level processing is too lacking in knowledge of what is being perceived to reliably continue the recognition process. It is at this point that higher level knowledge about the domain can be effectively utilized to put the lower level information into context. In object recognition systems, this information is usually contained in models that are used to relate the observables to the actual objects. The models are abstractions of the real physical objects that try to encode important information about the object in relation to the primitives and sensing environment being used. In some sense, the model information must be computable from the sensors. It is not enough to build descriptions of objects for realistic display; the models must contain criteria that are easily accessible to facilitate efficient matching of the model to a sensed object. The matcher is used to relate the two, and its job is facilitated by uncovering three dimensional structure through sensing.

3.1. UNDERSTANDING 3-D STRUCTURE

Object recognition in this work is predicated upon discovering three dimensional structure of objects which can then be matched against the models int he model data base.. It may seem obvious that understanding three dimensional structure is a necessary first step to a host of important robotic tasks, including recognition, grasping, manipulation and inspection. However, this has not been the primary approach of much previous work. Instead of being the primary initial focus, three dimensional structure was an outcome of the model matching phase. Only by correctly invoking a model (determined through a variety of viewpoint dependent and two dimensional projective analysis) was the actual three dimensional structure uncovered. By using active sensors, three dimensional structure can be discovered initially. The reasons why this is important are listed below:

- The sensed primitives need to be related to the model components in model based recognition. The models can be easily and efficiently structured as three dimensional surfaces and features. The discovery of three dimensional surfaces and features facilitates this matching effort. The models in this work use the same surface primitive that the sensors together compute. This eliminates expensive transformations of the data and possible information loss.
- Viewpoint independent recognition assumes no characteristic views of the object. The orientation in space of the object needs to be computed from the combination of sensing and high level reasoning. Uncovering the three dimensional structure makes this computation possible.
- There is a limit to the amount of recognition that can be done at the low level. Reasoning about three dimensional objects at a higher level implies understanding the three dimensional structure. Spatial relationships in three dimensions involve three dimensional entities. Only by uncovering these entities can higher level reasoning be invoked.
- Tasks beyond recognition also imply an understanding of three dimensional structure. Grasping, inspection and manipulation all involve understanding and reasoning about the three dimensional structure.

4. EXAMPLE: INTEGRATING SENSORS

The vision module consists of two CCD cameras that are calibrated with the robot workspace and registered for scan line coherence. The Marr-Hildreth edge operator [15] is applied to each of the images and zero-crossings of the convolved images are found. The zero-crossings are isolated to subpixels by a linear interpolation process to reduce the error due to quantization. These zero-crossings define homogeneous regions in the image from which region contours are extracted.

The matching phase uses the region contours as input. Isolated zero-crossings not on a contour are discarded, leaving sparse but stable contour match pixels. The matcher then attempts to match contour pixels using the constraints of scan line coherence and zero-crossing orientation and sign. The candidate match pixels are then correlated with regions of small window size centered on each candidate. Only those matches fulfilling the criteria above *and* having a correlation confidence level above 95% are accepted as match points. The outcome of this matching phase is a sparse set of match points on the contours of regions isolated from vision. As described in a previous paper there are limitations to the amount and accuracy of the data provided by the vision system. Stereo matching suffers from three main problems. The first is the inability of stereo to handle many candidate match points, such as is found in regularly textured objects. By using only sparse contour data the matcher becomes more accurate with few if any false matches. The second is the error due to quantization on a discrete pixel grid. For the camera geometry used here this can be 4 mm. The location of zero-crossings to subpixels reduces this error to 2 mm. The last problem is the inability of stereo to match horizontally oriented zero-crossings. There is no basis for distinction given the criteria above to choose between locally horizontal matches in a small region. Typically, zero-crossings whose orientation is more than 60° from vertical yield incorrect match results.

The outcome of stereo matching is shown in figure 3. There is sparse 3-D depth data on the contours, containing no horizontal matches. This is clearly not enough data to try to recreate surfaces and understand the object's structure. However, the data is accurate and reliable because it has been thinned and abstracted. It allows us to proceed to the next level of sensing with confidence, having sparse but accurate regions identified that can be used for further sensing. Attempts to drive the vision modules beyond this capability will invariably lead to a potentially serious error. The key idea is that *less is more* in the case of multiple sensing. We do not have to rely on this single modality for all our sensory inputs, only those it can *reliably* produce.

5. TACTILE SENSING

The vision module independently calculates regions of interest, giving limited 3-D contour information to the control module. The tactile module can now begin to sense the regions isolated from vision. A possible strategy is to have the tactile and vision modules work in parallel, increasing throughput. However, this approach ignores the fact that touch is an active sensor. Touch cannot succeed in a blind fashion. It needs control information to work reliably, and that control is provided by the touch. The tactile subsystem contains three levels. The top level is the control module in figure 2 that gives region information to the PUMA 560 arm running under VAL-II control. This level is responsible for the following functions:

- Orienting and positioning the tactile sensor to correctly approach an identified region.
- Determining whether a region isolated from vision is a surface, hole or cavity.

The third level is the tactile sensor which contains its own controlling Z-80 microprocessor. The tactile sensor is a finger shaped device that contains 133 pressure sensitive sites. This level is responsible for:

- Setting up signaling thresholds for the sensor.
- Converting the analog sensor signal to digital.
- Noise reduction and smoothing of the digitized signals
- Providing feedback of contact/nocontact to the arm control at the level above.

This hierarchy is used to orient and position the sensor and probe a region. Probed regions will reveal themselves to be surfaces, holes or cavities. A region is identified by the higher level sending its probable location and a surface normal approximation to the arm control level. This level then orients the arm and sensor accordingly, sets up parameters for the tactile subsystem, and moves the arm according to the feedback from the tactile sensor. The lowest level processes the sensor signals on its surface continuously, interrupting the level above (arm control) if contact occurs. If contact occurs the arm control level isolates the contact in space through its model of the sensor's geometry and reports back a surface to the top level controlling module. If the arm control monitors the distance the sensor travels without contact occurring, it reports back to the top level that a hole is present. If surface contact occurs after a distance D_{cav} has been traveled, then a cavity is reported. All three levels of the hierarchy are involved and communicate only what is necessary to the higher and lower levels to accomplish the task.

6. INTEGRATING VISION AND TOUCH

The initial determination of the regions structure (surface, hole, cavity) generates another round of sensing in the hierarchy to yield quantitative analysis of the regions to be used for the later matching phase. In the case of a surface, a surface patch needs to be interpolated from the sensory data. The vision data is too sparse to accomplish this, but the tactile sensor can trace the surface in an intelligent manner to build an accurate surface description. The method used is described in detail in [2]. The procedure is to build a Coons' patch representation [6] which is a particular form of bicubic surface patch used primarily in computer graphics and computer aided design. The patches are constructive in that they are built up from known data and are interpolants of sets of three dimensional data defined on a rectangular parametric mesh. This gives them the advantage of axis independence which is important in synthesizing these patches from sensory data. Being interpolating patches, they are able to be built from sparse data. The most important property possessed by these patches is their ability to from composite surfaces with C^2 (curvature continuous) continuity. The object domain (bowls, mugs, pitchers, plates) contains many curved surfaces which are difficult or impossible to represent using polygonal networks or quadric surfaces.

Starting with the contour data derived from vision, the tactile sensor is used to trace across each region, creating 4 new patches that are curvature continuous. The method is hierarchical in that each of these patches can then be subdivided by tactile tracing into a larger number of curvature continuous patches that more accurately interpolates the surface. Figure 4 shows how the method works, subdividing each contour into a set of knot points that create 4 boundary curves on a patch. The tactile sensor then traces across these patches, creating the new surfaces. Once again, the hierarchy is used as the top level control defines the boundary curves and the start and end points of the traces across the surface. These parameters are communicated to the lower level arm control which orients and positions the arm and begins its trace across the surface. The bottom level finger sensor control is used to generate contact feedback which the arm control level analyzes to plan a movement across the surface to stay in contact. The reported data from these contacts is communicated up to the control level which integrates the trace data into the format necessary for the Coons' patch representation. Figure 5 shows the interpolated surface patch for the pitcher's main body that is built from integrating real stereo data and active tactile tracing.

As can be seen from the above example, the hierarchical control works well. The important ideas of independent operation of each of each level, data abstraction and limited communication between levels, and developing strategies to maximize each sensor's most reliable mode of operation have been used. Further, the sensors work in a *complementary* mode whereby the passive vision data is used to guide the active tactile sensor.

A similar integration procedure is used to build hole and cavity descriptions which are useful in the matching phase of the recognition process. The tactile system is able to sense a hole or cavity's boundary and report this back to the control level where it also is used in the matching phase. Figure 6 shows a set of zero-crossings for one of the images of a coffee mug, the sparse stereo match points, and the interpolated surface of the front of the mug and the traced boundary curve of the hole. Both of these quantities are powerful matching entities, which allow determination of the object from the model data base as well as its orientation and position in the workspace.

7. SUMMARY

The integration of multiple sensors is important for more complex robotic tasks such as object recognition, grasping and manipulation. The integration tends to be task and sensor specific; however there are some general principles which work well in certain environments. In particular, the ideas of hierarchical control, data abstraction and compression between levels, and complementary sensing have been shown to be useful in an integrated system using passive vision and active touch sensing.

References

1. Agin, G., "Representation and description of curved objects," *Stanford University A.I. Memo*, no. 173, October 1972.
2. Allen, Peter, "Object recognition using vision and touch," Ph.D. Dissertation, University of Pennsylvania, Philadelphia, September 1985.
3. Bajcsy, R. and L. Lieberman, "Texture gradient as a depth cue," *Computer Graphics and Image Processing*, vol. 5, no. 1, pp. 52-67, March 1976.
4. Binford, T., "Survey of model based image analysis," *Int. Journal of Robotics Research*, vol. 1, no. 1, pp. 18-64, Spring 1982.
5. Bolles, R. C., P. Horaud, and M. J. Hannah, "3DPO: A three dimensional part orientation system," *Proc. 8th IJCAI*, Karlsruhe, Germany, August 1983.
6. Faux, I. D. and M. J. Pratt, *Computational geometry for design and manufacture*, John Wiley, New York, 1979.
7. Fisher, R. B., "Using surfaces and object models to recognize partially obscured objects," *Proc. IJCAI 83*, pp. 989-995, Karlsruhe, August 1983.
8. Grimson, W. E. L., *From images to surfaces: A computational study of the human early visual system*, MIT Press, Cambridge, 1981.
9. Grimson, W. E. L. and Tomas Lozano-Perez, "Model based recognition and localization from sparse three dimensional sensory data," A.I. memo 738, M.I.T. A.I. Laboratory, Cambridge, August 1983.
10. Hillis, W. D., "A high resolution imaging touch sensor," *Int. Journal of Robotics Research*, vol. 1, no. 2, pp. 33-44, Summer 1982.
11. Hoffman, D., "The interpretation of visual illusions," *Scientific American*, pp. 154-162, November 1983.
12. Horn, B. K. P., R. Woodham, and W. M. Silver, "Determining shape and reflectance using multiple images," *AI memo 490*, MIT AI Laboratory, Cambridge, 1978.
13. Kender, J. R., "Shape from texture: a brief overview and a new aggregation transform," *Proc. DARPA IU Workshop*, pp. 79-84, November 1978.
14. Marr, David and Tomaso Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, pp. 283-287, 1976.
15. Marr, David and Ellen Hildreth, "Theory of edge detection," *Proc. Royal Society of London Bulletin*, vol. 204, pp. 301-328, 1979.
16. Nitzan, D., "Assessment of robotic sensors," *Proc. 1st International Conference on Robot Vision and Sensory Controls*, Stratford-upon-Avon, UK, April 1-3, 1981.
17. Oshima, M. and Y. Shirai, "Object recognition using three dimensional information," *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 4, pp. 353-361, July 1983.
18. Overton, K. J., "The acquisition, processing and use of tactile sensor data in robot control," Ph.D. Dissertation, University of Massachusetts, Amherst, May 1984.

19. Shapiro, Linda, J. D. Moriarty, R. Haralick, and P. Mulgaonkar, "Matching three dimensional models," *Proc. of IEEE conference on pattern recognition and image processing*, pp. 534-541, Dallas, August 1981.

20. Shneier, M., S. Nagalia, J. Albus, and R. Haar, "Visual feedback for Robot Control," *IEEE Workshop on Industrial Applications of Industrial Vision*, pp. 232-236., May 1982.

21. Stevens, Kent, "The visual interpretation of surface contours," *Artificial Intelligence*, vol. 17, pp. 47-75, 1981.

22. Tomita, Fumiaki and Takeo Kanade, "A 3D vision system: Generating and matching shape descriptions in range images," *IEEE conference on Artificial Intelligence Applications*, pp. 186-191, Denver, December 5-7, 1984.

23. Witkin, Andrew, "Recovering surface shape and orientation from texture," *Artificial Intelligence*, vol. 17, pp. 17-47, 1981.

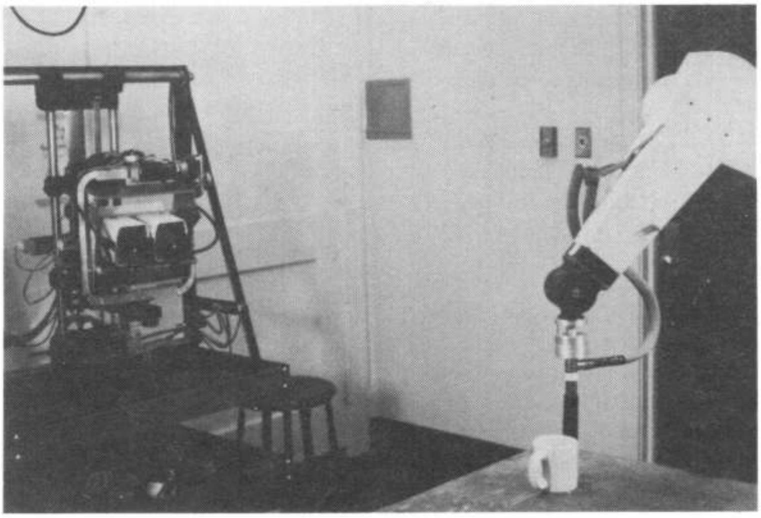

Figure 1. Experimental hardware.

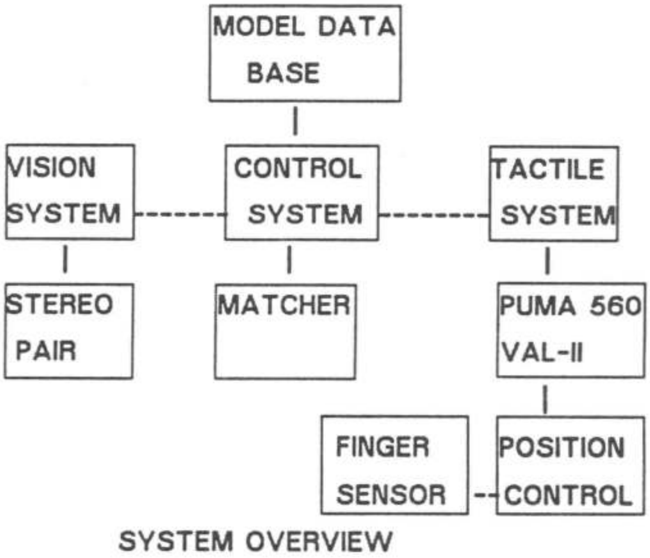

Figure 2. System overview.

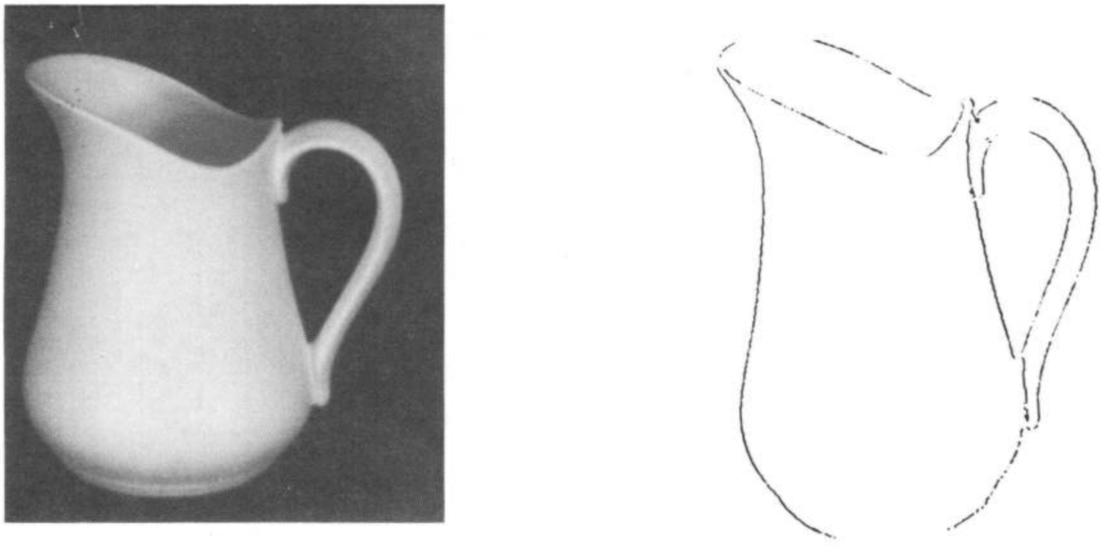

Figure 3. Original image and sparse stereo match points.

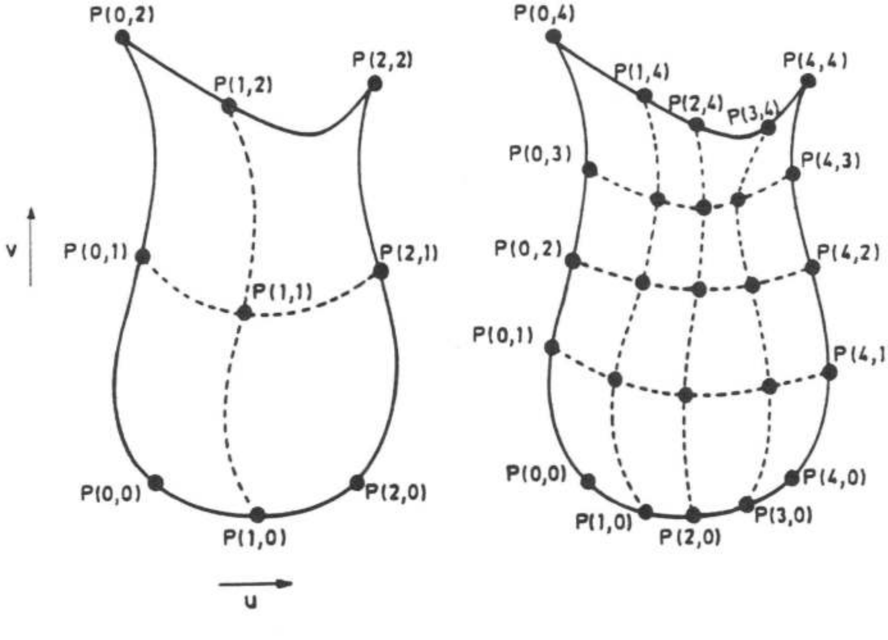

Figure 4. Creating surface patches by tactile tracing.

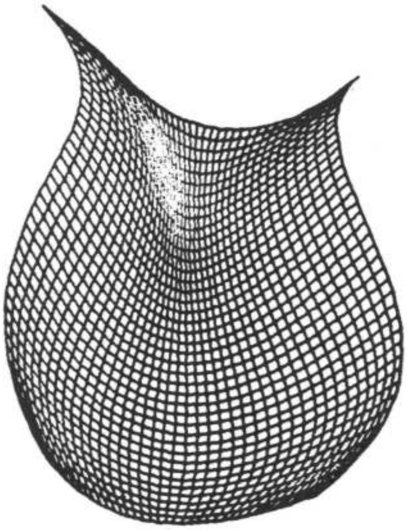

Figure 5. Interpolated surface patch .

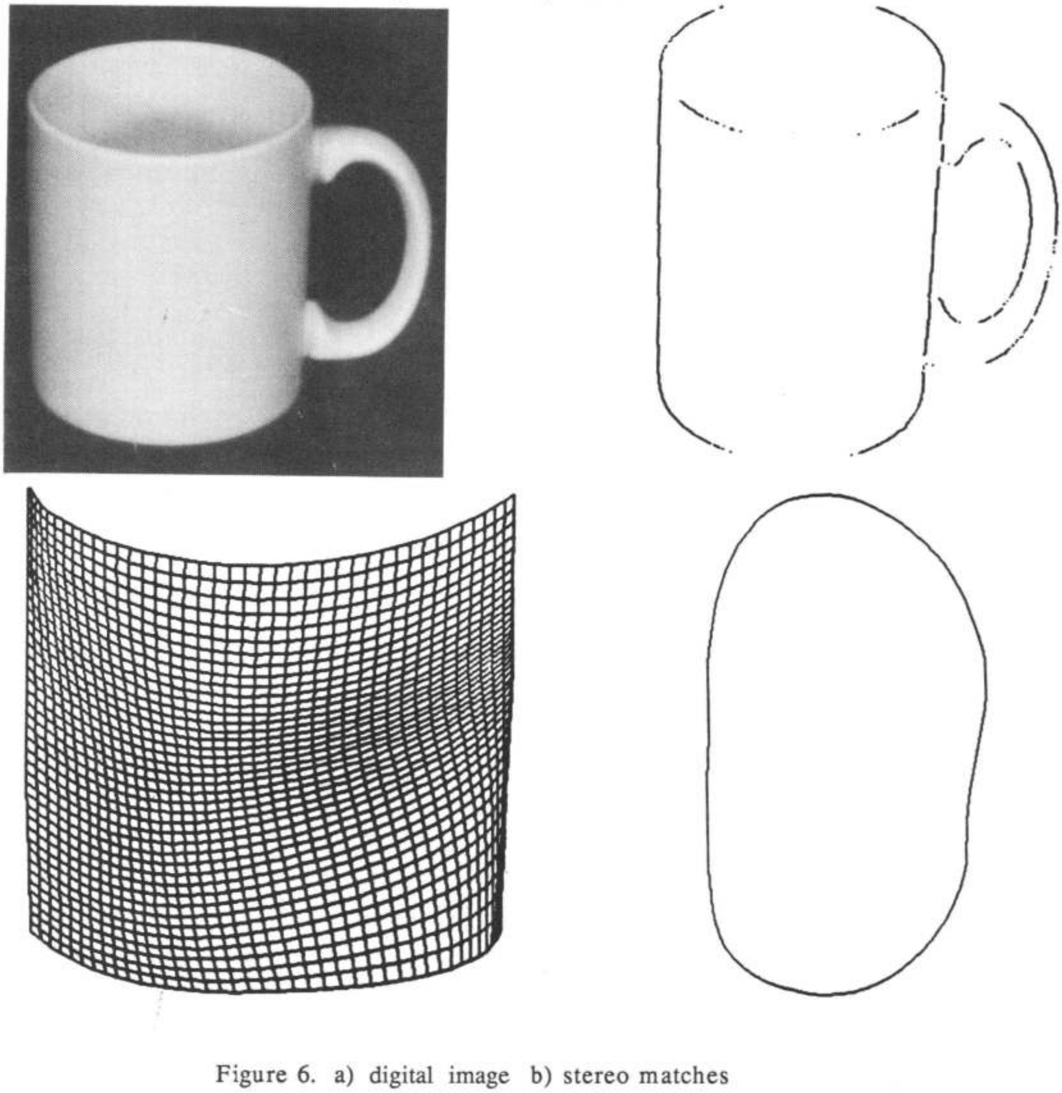

Figure 6. a) digital image b) stereo matches

c) interpolated surface d) traced boundary of hole